Inf Syst E-Bus Manage
DOI 10.1007/s10257-016-0329-4

ORIGINAL ARTICLE

Using association rules to assess purchase probability in online stores

Grażyna Suchacka[1] · **Grzegorz Chodak**[2]

Received: 8 March 2016 / Revised: 6 July 2016 / Accepted: 26 August 2016

Abstract The paper addresses the problem of e-customer behavior characterization based on Web server log data. We describe user sessions with the number of session features and aim to identify the features indicating a high probability of making a purchase for two customer groups: traditional customers and innovative customers. We discuss our approach aimed at assessing a purchase probability in a user session depending on categories of viewed products and session features. We apply association rule mining to real online bookstore data. The results show differences in factors indicating a high purchase probability in session for both customer types. The discovered association rules allow us to formulate some predictions for the online store, e.g. that a logged user who has viewed only traditional, printed books, has been staying in the store from 10 to 25 min, and has opened between 30 and 75 pages, will decide to confirm a purchase with the probability of more than 92 %.

Keywords Association rules · Data mining · Web usage mining · Click-stream analysis · Log file analysis · e-Commerce

1 Introduction

Along with the development of the information society and the knowledge-based economy, electronic commerce has been gaining increasing popularity all over the world. The most common type of e-commerce has been B2C (*Business-to-Consumer*) trade, typically realized through online stores. In an electronic

✉ Grażyna Suchacka
gsuchacka@uni.opole.pl

[1] Institute of Mathematics and Informatics, Opole University, ul. Oleska 48, 45-052 Opole, Poland

[2] Department of Operations Research, Wrocław University of Technology, Wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław, Poland

Published online: 06 September 2016

environment, detailed data on e-customer behavior may be easily collected and analyzed using data mining techniques. The acquired knowledge may then be used to improve the customer service in a Web store, to increase customer satisfaction, and consequently, to raise the online store's conversion rate in the long run.

In this paper we consider a typical online store implemented as a B2C website hosted on a Web server supported by application and database servers. The website is composed of many pages, on which a user may perform various functions, such as browsing the store's offer, adding selected items to a virtual shopping cart, or confirming an order (which means actually making a purchase online). The site may be simultaneously accessed by many Web users via their Internet browsers. A user interacts with the site through a series of related page requests made during a single *user session*.

We propose discovering associations between various features of user sessions in respect of purchases realized in a Web store (an online bookstore in our case). We distinguish two groups of customers: *traditional customers* who view and sometimes also buy only typical products (printed books) and *innovative customers* who view and sometimes buy not only typical products but also innovative ones (audio-books and multimedia products). We aim at identifying these sessions' features for both customer groups which indicate a high probability of making a purchase in session. To this end we apply association rule mining.

The motivation for our study was the problem of unpredictability of customers' behavior in online stores and the need for online retailers to be capable of identifying potential or future buyers based on their online behavior. In practice, only a small percentage of visitors in online stores become buyers. Thus, the online retailer's ability to predict buying sessions increases chances of achieving a competitive advantage by focusing on a group or groups of key customers.

This ability is even more important in the face of unpredictability and high variability of Web traffic. Occasional peak times may result in the server overload and consequently, in long delays, abandoned shopping carts, and incomplete transactions. Such situations may be often observed on Black Friday and Cyber Monday, the busiest online shopping days of the year, following Thanksgiving Day in the United States. The ability of online retailers to predict purchases would make it possible to implement an intelligent, business-oriented admission control and scheduling policy on a Web server, e.g. in such a way that under the heavy load a website could offer a higher quality of service to visitors with a higher probability of making a purchase.

Focusing on specific customer groups was motivated by results of some previous studies, e.g. Chang et al. (2007) and (Shim et al. (2012), showing that it is worth focusing on key customers of an online store instead of all visitors. Distinguishing *innovative* and *traditional* customer groups resulted from the experience of the online bookseller and the fact that the vast majority of transactions in the bookstore is finalized just by these two groups of customers.

Our paper is organized as follows. In Sect. 2 we present a format of a typical Web server log file and discuss an issue of the click-stream analysis based on log data. In Sect. 3 we review related work on e-customer behavior analysis in respect of discovering association rules in online stores' data and predicting customers'

purchase intentions. In Sect. 4 we discuss our research methodology and formulate the problem of discovering association rules taking user session features into account. In Sect. 5 we present results of the analysis performed for a real online bookstore data for two customer groups and discuss the results in Sect. 6. Section 7 concludes the paper and indicates outlook for further research.

2 Web server log file analysis

Web user behavior has been typically analyzed using data recorded in Web server access logs. Logs are text files in which some data on all HTTP requests coming to the server from Web clients are automatically recorded. For example, in the NCSA Combined log format the following data for each request are written:

- IP address or host/subdomain name of the HTTP client (i.e. the client Web browser),
- identifier of the Web client (optional field, usually not given),
- username or user id used for authentication (optional field, usually not given),
- date and time stamp of the HTTP request,
- HTTP method (the most popular is GET, corresponding to downloading data from the server),
- version of HTTP protocol (HTTP/1.0 or HTTP/1.1),
- URI (*Uniform Resource Identifier*) of the requested server resource,
- HTTP status code (e.g. 200 means that the request was successfully processed at the server),
- the number of bytes of data transferred for the HTTP request in response,
- referrer, i.e. the URL (*Uniform Resource Locator*) which linked the user to the site (optional field),
- user agent string describing the client Web browser (optional field).

The initial motivation for log maintenance was to collect information useful for troubleshooting possible server errors. However, it has turned out with time that logs can be the source of valuable information on the server system load and server performance, as well as Web user navigational and transactional patterns. Some of this information may be easily acquired by using simple statistical tools but other require applying more sophisticated analytical techniques. Many basic problems are connected with log file analysis, the main ones being very large data size and too many analytical methodologies (Sen et al. 2006).

The analysis of user behavior on a website is referred to as a click-stream analysis. Click-stream data corresponds to series of Web pages requested by users within their sessions and it may be reconstructed from logs. Such analyses have a very high practical value to online retailers as they make it possible to understand the way in which customers use the site and navigate through the store, especially in the context of successful purchase transactions. The proper analysis may lead to better organization of an e-commerce service and more efficient business decisions. Adnan et al. (2011) emphasized that the goal of log file analysis should be gaining

an insight into the user browsing behavior and then translating this insight knowledge into foresight knowledge that would help the online seller to adjust their business policies. Better understanding of e-customer shopping behavior makes it possible to reduce customers' product searching time and thus to decrease searching cost by recommending products customers may probably be interested in.

The click-stream analysis based on log data involves many problems. A critical step is data pre-processing which includes user identification, session identification, path completion, and transaction identification (Huiying and Wei 2004). Two of the biggest impediments to collecting reliable usage data include local caching and proxy servers. In a server log all requests from a proxy server have the same identifier although the requests may represent many users. Furthermore, due to proxy server level caching a single response from the Web server can be viewed by multiple users throughout some period of time, thus distorting the image of user sessions at the Web server (Cooley et al. 1999). One of the unavoidable data preparation tasks necessary for the success of log mining includes the proper identification of user sessions. Two main approaches may be applied here (Chen et al. 2004): interval sessions, where each session consists of pages accessed by the same user within a time limit and gap sessions, where each session is a sequence of pages accessed by the same user with pairwise access time gaps below a threshold value. A 30-min threshold has been typically assumed (Adnan et al. 2011; Chen et al. 2004; Catledge and Pitkow 1995; Stevanovic et al. 2011; Suchacka and Chodak 2013). Another important preparation task before the click-stream analysis is identification and elimination of traffic generated by bots which reveal different navigational patterns than human users (Suchacka 2014; Stassopoulou and Dikaiakos 2009).

Many studies on the click-stream analysis have addressed the problem of user navigation paths and sequential patterns discovery on e-commerce sites (Adnan et al. 2011; Kwan et al. 2005; Lee and Yen 2007; Shim et al. 2012). A key observation has been connected with the fact that visitors to an online store are potential buyers and perform different kinds of operations on different pages. Depending on an operation, each page may be assigned to some *session state* – typical states on a B2C site can be Home page, Login, Register, Browse, Search, Select, Add to shopping cart, and Pay. After distinguishing session states for a given e-commerce website a user session model may be developed for all customers (Jenamani et al. 2003; Kwan et al. 2005; Lee et al. 2001) or for different customer groups (Chang et al. 2007; Kim and Cho 2003; Nenava and Choudhary 2013; Shim et al. 2012; Wang et al. 2004). User session models have been used to develop request admission control and scheduling algorithms for Web servers in order to improve the quality of service on e-commerce websites (Borzemski and Suchacka 2010; Suchacka and Borzemski 2013; Totok and Karamcheti 2006; Zatwarnicki and Zatwarnicka 2014; Zhou et al. 2006). Furthermore, many research studies have applied various data mining techniques to analyze, support, and predict user behavior (Borzemski and Kamińska-Chuchmała 2012; Chen et al. 2009; Cheng and Chen 2009; Huk et al. 2015; Mohammadnezhad and Mahdavi 2012; Poggi et al. 2007; Shen and Su 2007; Suchacka et al. 2015b; Van den Poel and Buckinx 2005;

Wrzuszczak-Noga and Borzemski 2013). In the next section we classify and review approaches which applied association rules to online store data.

3 Related work

Associations reflect relationships between individual objects and are characterized by some strength measures reflecting their quality or significance (Kazienko 2008). *Association rules* are very popular types of associations explored on the Web and in electronic commerce. Discovering association rules (also called market basket analysis or affinity analysis) involves finding mutual, often hidden, associations between attributes characterizing objects in a data set. An association rule allows one to describe quantitatively the relationships between the attributes, and has the form: "if the antecedent, then the consequent" (where the antecedent and the consequent are sets of attributes) along with measures of confidence and support of the rule. The example rule determined for an online store could be the following: "60 % of customers who buy glasses, buy a spectacles case as well, with 30 % of all customers purchasing both products at the same time". In this rule, the antecedent contains one attribute—the event "the customer bought glasses", the consequent contains an attribute corresponding to the event "the customer bought a spectacles case", 60 % means the rule confidence, and 30 % means the rule support.

In general, attributes of the rule antecedent and consequent are not necessarily limited to products in a physical sense but may be based on any event. In the case of user sessions in online stores the attributes under consideration have included products purchased together, Web pages visited together, customer navigation paths, customer requirements expressed by phrases used, and user session features. In the following Subsections we review related work on discovering association rules in online stores' data to show the applicability of this data mining technique in various areas of e-commerce.

3.1 Discovery of associations between products purchased together by different customers

The first and most common application of association rules has been the analysis of large databases of customer transactions to discover relations between products purchased together by different independent customers. The overall goal of the analysis has been sales support. In (Agrawal et al. 1993) authors proposed an efficient algorithm for generating significant association rules between sets of purchased items. The algorithm uses two novel estimation and pruning techniques to avoid measuring certain item sets while guaranteeing completeness. It also incorporates buffer management to cope with the potential problem of insufficient memory during measuring the huge number of item sets (transactions). Application of the algorithm to sales data obtained from a large retailing company showed its high effectiveness.

Many improved algorithms for finding association rules in large item sets have then been proposed, e.g. SQL query-based SETM algorithm (Houtsma and Swami

1995), AprioriHybrid algorithm (Agrawal and Srikant 1994), algorithm DHP (*Direct Hashing and Pruning*) (Park et al 1997), algorithm FP-Growth (*Frequent Pattern-Growth*) (Han et al. 2004), the multiple minimum supports mining algorithm using maximum constraints (Lee et al. 2005), algorithm CBAR (*Cluster-Based Association Rule*) (Tsay and Chiang 2005), DI-Apriori algorithm for mining dissociation rules (Morzy 2006), or algorithm MIbARM (*Matrix and Interestingness-based Association Rule Mining*) (Deng et al. 2010).

Peng and Wan (2010) proposed mining association rules between products based on rough sets. From different types of products different feature vectors may be extracted. By using a discernibility matrix and discernibility function a reduction set of products may be obtained which is then used to derive association rules. The resulting set of rules is then verified by applying the correlation analysis.

Thuan et al. (2012) proposed extension of association rules between products purchased in an online store with a temporal dimension. The authors investigate cyclic (temporal) association rules during time intervals which follow some user-given time schemas. An example of a time schema is (day, month, year) where the notation (1, *, 2011) means the time interval consisting of all the 1st days of all months in year 2011. Cyclic association rules should have the minimum confidence and support at regular time intervals and do not have to be in force for the entire transactional database but rather only for transactional data in particular periodic time intervals. The authors discuss the A-priori-based algorithm MTP (*Mining of Time Pattern association rules*) and verify its performance with experiments performed on a real sales database.

It was observed that different users follow different navigation paths on a B2C site and request different pages in different ways and with various frequencies. Taking this observation into consideration, some studies have applied clustering methods for creating customer profiles and explored associations between products purchased together by customers in individual clusters, e.g., (Kwan et al. 2005; Nenava and Choudhary 2013; Mohammadnezhad and Mahdavi 2012; Tanna and Ghodasara 2012), or by key customers (Chang et al. 2007; Shim et al. 2012). Results have been used to provide customers with personalized recommendations, to optimize a website structure, and to refine a CRM (*Customer Relationship Management*) strategy.

Mohammadnezhad and Mahdavi (2012) presented a new model for a tourism recommendation system which proposes tours to visitors using two data mining techniques: clustering and association rules. According to the model, customers are initially clustered using SOM (*Self Organize Map*) algorithm to determine the number of clusters and *k*-means algorithm to generate clusters. Then, by analyzing tours ordered by the customers in the past, association rules are created separately for each cluster using A-priori algorithm. Recommendations for an active visitor on the site are made based on the cluster in which the targeted tourist is and their past shopping history.

A similar idea, also combining customer clustering and discovery of associations for customer clusters, was under the approach discussed in (Tanna and Ghodasara 2012). In this study the vector quantization-based clustering was performed to categorize e-customers based on their RFM values and then the A-priori association

rule mining algorithm was applied to find out relationships among customer purchases in individual clusters.

In (Chang et al. 2007) an anticipation model of potential customers' near future purchasing intentions has been proposed. The model is inferred from past purchasing behavior of loyal customers and log file data for loyal and potential customers using clustering analysis and association rules. The basic assumption is that customers always log in to the online store. The first step is a selection of star products that contribute to a large percent of company sales (the model uses only one star product as the input data so it should be run separately for each selected star product). The second step is establishing loyal customer profiles. From a set of all customers who purchased the star product a group of loyal customers is extracted by calculating each buyer's PPT (*Past Purchasing Tendency*) value. Loyal customers are characterized with some personal backgrounds, including age, gender, education level, family size, etc. Clustering analysis of loyal customers is performed based on their PPT values and personal backgrounds in order to create loyal customers' profiles and infer their past purchasing tendency of each characteristic of personal information. The third step is searching for potential customers by comparing loyal customer profiles to the personal information of customers who have never purchased the star product before. For each potential customer their recent intentions are measured by analyzing their recent Web log data via association rules and a near-future purchasing probability is determined. Potential customers whose recent behavior match the association rules are most likely to buy a star product so an online recommendation system should provide them with information about that product.

In (Cho et al. 2013) an algorithm IWMAR (*Incremental Weighted Mining Association Rules*) using FP-tree was proposed for an online store. Weighted association rules were applied to find relations between products taking into consideration information on product purchasability. Segmentation by product usage is performed with RFM (*Recency, Frequency, Monetary value*) analysis and weights in the rules are based on the product RFM scores. The generated rules are used in an online recommendation system to predict and recommend items with high purchasability.

Shim et al. (2012) discovered association rules and sequential patterns by analyzing the transaction data of an online shopping mall. After defining VIP customers in terms of recency, frequency, and monetary (RFM) value, the authors developed a model classifying customers into VIP or non-VIP using such data mining techniques as artificial neural network, decision tree, logistic regression and bagging. Then, for a VIP group association rules and sequential patterns among product categories and subcategories were discovered. Research was based on a dataset consisting of five tables: demographic, bulletin, comment, order management, and product order table. Twelve input variables were taken into account for the classification of the customers into VIP or non-VIP: all three RFM values, age, channel of registration (search or recommendation), and other variables connected with the customer preferences in paying, product delivery, activity on the bulletin board and in commenting on products. Results of the analysis were used to suggest elements of CRM strategy for the online shopping mall: the classification model

may be used to identify key customers to which marketing activities should be exercised; the website should be redesigned so that pages associated with highly associated product categories should be one-click away from each other; a set of keywords of the online shopping mall should include or imply these product categories which are strongly included in most discovered association rules and sequential patterns.

3.2 Discovery of associations between products purchased together combined with information about customers' navigation paths

In (Lee and Yen 2007) an algorithm IWA (*Integrating Web traversal patterns and Association rules*) was developed for mining Web transaction patterns in an online store. The proposed algorithm takes both the navigation and purchasing behaviors of e-customers into consideration at the same time, i.e. it combines the knowledge on the sequences of pages visited in buying sessions and types of products added to shopping carts on particular pages having occurred in these sequences. Furthermore, the algorithm uses information about a website structure to prune unnecessary candidates for association rules and it considers both user forward traversal information and backward information. The goal of the approach is to optimize a website structure to increase the e-commerce conversion rate.

Kwan et al. (2005) developed the eCB (*e-Customer Behavior*) model which uses association rules and online analytical mining techniques for continuous discovery of e-customer click and tick sequences based on data in Web server logs and cookie files. They proposed a mental cognitive model which provides a basis for describing e-customer behavior in four dimensions: access frequency, revisit recentness, session duration, and path length. The movement of users in an online store was modeled by an e-customer behavior graph in which a generic navigation path on a site was divided into three phases: e-customer awareness, exploration, and commitment. The proposed eCB model discovers knowledge about contiguous associations between the awareness and exploration, and between exploration and commitment in order to identify the critical pages that allow users to progress from phase to phase until the purchase confirmation. The general goal of the model is e-customer segmentation into different profiles, determination of critical pages affecting the click direction and sequence, and discovery of individual customers' profiles. Such knowledge may allow the redesign of a B2C site to incur positive voluntary clicks from visitors by directing them to the purchase commitment phase.

3.3 Discovery of associations between customer requirements represented by a set of phrases and products purchased by the customer

Zhang and Jiao (2007) addresses the problem of product recommendation in online stores. All products available in a store are described with class labels. By semantic analysis of a historical transaction database, customer requirements are described as a set of phrases. Each purchase transaction is represented as a pair of a set of phrases and a class label of the purchased product. The generated association rules have sets of phrases in a rule antecedent and class labels in a rule consequent. Thus,

generation of recommendations for a new customer consists in inferring class labels from requirements of that customer. The feasibility of the proposed recommendation system was successfully validated with a prototype for personalization in mobile phone B2C e-commerce applications.

Nenava and Choudhary (2013) proposed a hybrid recommendation system in m-commerce based on innovative *k*-means clustering and DAR (*Distributed Association Rules*). For a given website firstly, clustering is performed to distinguish groups of similar users based on their searches and to create user profiles. To this end, the quantity of independent searches made by all users in the similar group is utilized. Unique searches made by users in a group and the searched parameters are identified and frequency of all comparable searches of each user is originated. Then two techniques are used to create group user profiles. In the initial technique, matrix-based methods are applied; for each cluster top rated searched objects are determined using popular two-dimensional reduction techniques (DAR, NNMF, PCA, SVD) and saved in a user profile. In the next technique, vectors of searches for every user are generated; then the vectors are clustered to discover similar users and the searches made by similar users are used to discover associations. Lastly, product features which can differentiate between two searches are extracted by these associations. The mechanism used for rule generation should be able to clearly recognize concrete products in the database (e.g. make and model in the case of mobiles). As a result, a set of unique high-confidence association rules is assigned to each group user profile and these top rules can then be used as recommendations for active users belonging to the respective cluster.

3.4 Discovery of mutual associations between Web pages visited together in session

Adnan et al. (2011) developed an integrated approach to analyze Web log data that involves statistical analysis, sequential association rules discovery, and social network construction and analysis. After determining frequent sequences of visited pages for an example site, association rule mining was performed on the sequences in order to determine correlations between the page sequences frequently visited by users. Furthermore, the links between pages were analyzed by constructing a social network based on the frequency of access to the pages in such a way that two pages get linked in the social network if they are identified as frequently accessed together. For the generated network centrality measures, highly ranked pages, islands, hubs, and authorities were determined. The results may be helpful in guiding e-commerce business promotions, i.e. to provide the online retailer with the information needed to dynamically display targeted advertisements on respective Web pages to their e-customers: what, where, and when to promote.

3.5 Discovery of associations between e-customer values and a market type

Besides the research aimed at improving customer service in online stores there have been studies considering the application of association rules in CRM for e-commerce. For example, Chiang (2011) proposed a methodology to mine

association rules of customer values. The research was based on data collected by questionnaires from experienced online shoppers within 6 months in Taiwan. Each shopper was described with five variables of consumer values according to an improved RFMDR model: R (recency), F (frequency), M (monetary), D (discount, a variable reflecting a benefit contribution), and R (returning shipping cost, a variable corresponding to the cost of sales). Using Ward's method, online shoppers were partitioned into three market types: practical oriented, additional cost, and enjoy shopping. These three clusters were positively verified by discrimination analysis. Then, associations between five RFMDR model variables and a market type were discovered using A-priori algorithm. The established rules may be used to refine a CRM strategy, e.g. online stores may use different marketing projects for customers with the same customer values but belonging to two different markets. Other applications of association rules in CRM domain aimed at improving one to one marketing (Adomavicius and Tuzhilin 2001), online personalized sales promotion (Changchien et al. 2004), or customer churn prediction (Tsai and Chen 2010).

3.6 Discovery of mutual associations between user session features

Markov and Larose (2007) applied A-priori algorithm to discover mutual associations between various user session features. To this end, they explored association rules with attributes based on user session features contained both in the rule antecedent and subsequent. Similarly, Carmona et al. (2012) combined such data mining techniques as clustering, association rule learning, and subgroup discovery to suggest improvements of a concrete e-commerce website design. Clustering of user sessions was performed through *k*-means algorithm and association rules were generated using A-priori algorithm. Then, subgroups were discovered using the evolutionary fuzzy algorithm NMEEF-SD (*Non-dominated Multiobjective Evolutionary algorithm for Extracting Fuzzy rules in Subgroup Discovery*). The authors took into consideration different variables describing user sessions: the type of Web browser used, the referrer, keywords, the visitor type (unknown or regular), session duration, the number of page views, and the number of unique page views. The analysis of mutual associations between these features was performed, giving some rules interesting for a website designer, e.g.: users who access the website directly, remain on the website during an accepted time; reference websites with external links to the analyzed online store result in visits with a low number of page views and unique page views; the majority of users who visit the site through a search engine, use the Internet Explorer browser and visit between one and one hundred pages. Based on the discovered rules, guidelines for improving the site usability and user satisfaction were suggested.

In our previous study (Suchacka and Chodak 2013) user session features have been used to build association rules as well. We proposed discovering associations between e-customer session features and a buying action in session and presented preliminary results of our research. In this paper, we expand the set of session features with the source of the customer visit and key products' categories. Moreover, we distinguish two customer groups based on viewed product categories

and identify these session features which indicate the high probability of making a purchase in session of each group.

To sum up, the analysis of related work shows that although the idea of discovering association rules to analyze e-commerce transactions is not new, none of the previous studies aimed at assessing a purchase probability in an online store using association rules. The vast majority of studies explored associations between products purchased together by different customers, mainly for the purposes of online product recommendation, e.g., (Chang et al. 2007; Cho et al. 2013; Mohammadnezhad and Mahdavi 2012; Peng and Wan 2010; Shim et al. 2012; Thuan et al. 2012; Tanna and Ghodasara 2012). Other studies explored associations between customers' navigation paths and products purchased by the customers for the purposes of website structure optimization (Kwan et al. 2005; Lee and Yen 2007), associations between phrases used by customers and products purchased by them for the purposes of product recommendation system and website structure optimization (Nenava and Choudhary 2013; Zhang and Jiao 2007), mutual associations between web pages visited in sessions for the purposes of guiding business promotions and website structure optimization (Adnan et al. 2011), as well as mutual associations between various session features for the purposes of website structure optimization (Carmona et al. 2012; Markov and Larose 2007). Table 1 summarizes previous approaches to association rule discovery in e-commerce websites.

3.7 Methods for predicting purchase intention on B2C e-commerce sites

Analyses of Web user behavior on e-commerce sites have shown that the behavior of users deciding to buy online differs from the behavior of users who stay only visitors. As a consequence, some studies aimed at assessing the probability of making a purchase using historical data on customers' behavior. Various data mining techniques have been proposed to predict online purchases.

In (Van den Poel and Buckinx 2005) the contribution of different types of predictors to the purchasing behavior of e-customers was investigated. Logit modeling was applied to predict whether a purchase is made during the next customer's visit and Furnival and Wilson's global score search algorithm was used to find the best subset of predictors. An extensive set of variables from a few categories was taken into account, including such categories as general and detailed click-stream measures, past purchase behavior, and customer demographics. The best predictors of online purchasing behavior were detailed click-stream variables, however predictors from other categories proved to be important as well.

In (Poggi et al. 2007) an approach for identifying users with high purchasing intentions and assigning priorities to their sessions was proposed. The approach is based on traditional machine learning techniques and second order Markov-chain models. It uses both static and dynamic information on user sessions. The static information is obtained from historical data in Web server logs and includes: access time, session length, and the information whether a user logged into the site, whether the user was a returning customer, and whether the user has already bought in the past. The dynamic information is connected with the navigation path followed

Table 1 Comparison of approaches to association rule discovery in e-commerce websites

Study	Associations between…	Goal of the approach	Data mining techniques	Data source	All sessions/ clusters/a group
Kwan et al. (2005)	Customer navigation paths and products purchased by the customers	Website structure optimization	Association rules discovery Clustering	Server logs and cookie files	Clusters
Chang et al. (2007)	Products	Product recommendation (only for logged customers)	Association rules discovery Clustering	Transaction database	Group of loyal customers
Lee and Yen (2007)	Customer navigation paths and products purchased by the customers	Website structure optimization	Sequential association rules discovery	Server logs	All
Markov and Larose (2007)	Session features	Website structure optimization	Association rules discovery	Server logs	All
Zhang and Jiao (2007)	Phrases used by customers and products purchased by them	Product recommendation	Association rules discovery	Transaction database	All
Peng and Wan (2010)	Products	Product recommendation	Association rules discovery Rough sets	n/a	All
Adnan et al. (2011)	Web pages visited in session	Guiding business promotions, website structure optimization	Sequential association rules discovery Social network analysis	Server logs	All
Carmona et al. (2012)	Session features	Website structure optimization	Association rules discovery Clustering (k-means algorithm) Subgroup discovery (fuzzy rules extracting)	Google Analytics data	All
Mohammadnezhad and Mahdavi (2012)	Products	Product recommendation	Association rules discovery Clustering (SOM, k-means algorithm)	Transaction database	Clusters

Table 1 continued

Study	Associations between…	Goal of the approach	Data mining techniques	Data source	All sessions/ clusters/a group
Shim et al. (2012)	Products	Refining a CRM strategy, website structure optimization	Association rules discovery Sequential patterns discovery Classification (neural network, decision tree, logistic regression, bagging)	Transaction database	Group of VIP customers
Tanna and Ghodasara (2012)	Products	Product recommendation	Association rules discovery Clustering (vector quantization)	Server logs	Clusters
Thuan et al. (2012)	Products	Product recommendation	Temporal association rules discovery	Transaction database	All
Cho et al. (2013)	Products	Product recommendation	Weighted association rules discovery Frequent patterns discovery (FP-tree)	Transaction database	All
Nenava and Choudhary (2013)	Phrases used by customers and products purchased by them	Product recommendation, website structure optimization	Distributed association rules discovery Clustering (*k*-means algorithm)	Server logs	Clusters

by the user in their current session confronted with behavioral graphs based on a Markov-chain model. Two models were created based on the static information: one for purchasing users and another for non-purchasing ones. The machine learning techniques applied to user classification were a logistic linear regression, a decision tree, and a Naïve Bayes classifier.

Chou et al. (2010) proposed a Web usage mining approach to assess user's knowledge and needs for specific products and to predict user's intention of seeking, buying, and abandoning. The sequence mining technique was applied to analyze user's navigation paths on the e-commerce site and to discover navigation patterns. Then an Artificial Neural Network (ANN) was built to discover customer profiles according to the discovered patterns. The goal was classification of e-customers into groups with similar behavioral patterns and the ability to predict future customers' behavior connected with their buying intention or interesting in specific products.

Hop (2013) applied a few methods to predict a probability of placing an order by a user currently visiting a Web store: random forests, Support Vector Machines (SVM), and Artificial Neural Networks. Predictor variables were based both on behavioral data, like time or duration of a user session, and on detailed customer and transaction data, connected with products viewed, added to the shopping cart, and bought, product prices, customer lifetime values, number of past customer purchases, customer age and gender. The best prediction results were achieved by the random forest classifier, followed by the SVM classifier. The neural network classifier did not perform adequately.

A Support Vector Machine was applied to online purchase predictions in (Suchacka et al. 2015b) as well. User sessions in a Web store were reconstructed from data in server logs and each session was represented as a vector of session features. The problem of predicting purchases in a Web store was formulated as a supervised classification problem with two target classes: a buying session and a non-buying session. The authors built four SVM classification models with various kernel functions: radial, linear, polynomial, and sigmoid. The results showed a differentiated efficiency of the classifiers depending on the kernel. All four classifiers achieved a very high accuracy, i.e. the overall percentage of correct classifications for buying and non-buying sessions. However, the classifiers based on the radial and sigmoid kernel functions were ineffective in predicting buying sessions. The best performing classification model was the SVM with a linear kernel function.

A similar classification problem was considered in (Suchacka et al. 2015a), where the *k*-Nearest Neighbors (*k*-NN) method was applied to predict buying sessions in a Web store. Based on historical log data a *k*-NN classifier was built and its efficiency was evaluated for different neighborhood sizes. A distance measure between user sessions was the Euclidean metric. Evaluation of the approach showed that all the classifiers have a very high predictive accuracy. The classifier taking eleven nearest neighbors into account was the most effective both in predicting purchases in current user sessions and in terms of overall predictions.

Analysis of the literature showed that other data mining techniques may be successful in predicting online purchases. To the best of our knowledge, association rules have not been applied for this purpose so far. The advantage of association rule

discovery over other data mining techniques is that this method is relatively simple and does not require building and training a classification or prediction model. This advantage is especially significant in real-time applications like online stores. Furthermore, Web traffic characteristics and e-customer behavioral patterns may evolve with time so it is important for prediction methods implemented on e-commerce sites to be capable of adapting to traffic changes in real time. Other methods proposed for predicting purchase intention in online stores so far, like Support Vector Machines, Neural Networks, or the *k*-Nearest Neighbors classifier, require updating methods' parameters through the training procedure, which is algorithmically complex and very time-consuming. In contrast, association rule discovery according to our approach requires only continuous collection of data and performing their statistical analysis periodically, which may be performed online on a regular basis.

Our main contribution includes an approach applying association rules to online purchase prediction and verification of its efficiency with real e-commerce data for two key customer groups, distinguished based on viewed products' categories. Moreover, we proposed new session features to be used in association rule mining: mean time per page, type of the source of the visit, categories of products viewed in session, and the fact of performing key operations related to the purchase transaction: user's registration or logging into the site, adding a product to the shopping cart and confirming an order. Taking into consideration the last session feature allowed us to build association rules with this feature in the rule consequent. We are not interested in what product a customer is likely to buy in session (as it was considered in most of related work) but what is the probability of buying in a given session – this makes it possible to distinguish more valuable user sessions among many other active sessions on the online store website and to use this information in the process of session management and in taking business decisions.

4 Research methodology

4.1 Data collection

The analyzed website was an online bookstore built on an osCommerce platform, hosted on an Apache HTTP server on Linux with PHP and MySQL support. Web server logs had been configured according to NCSA Combined log format.

4.2 Reconstruction of user sessions from log file data

A dedicated C++ program was used to read data from logs, pre-process and clean it, reconstruct user sessions and perform the analysis. First, data corresponding to requests for embedded Web objects, such as image or video files, was eliminated. Thus, only requests corresponding to page requests (user clicks) were left. Based on them user sessions were reconstructed in the following way. Each unique user identifier was assigned to the corresponding sequence of page requests based on two requests' fields: the IP address and the user agent string. Then for each user

identifier successive sessions were distinguished using the gap session approach with a minimum 30-min threshold between two subsequent sessions of a given user.

Not all sessions reconstructed from log data are useful for the analysis of customer behavior. After the preliminary analysis the following sessions were excluded from further analysis:

- sessions performed by Web bots (search engine crawlers, shopping bots, etc.) which were identified using approach proposed in (Suchacka 2014),
- sessions performed by the website administrator or the administrative software,
- sessions containing less than two pages or lasting less than 2 s (such sessions may be considered "accidental" and they correspond to users who left the site just after entering it; such sessions are not useful for our analysis).

4.3 Characterization and refinement of user sessions

Each user session was characterized with the following session features:

1. session length (L), measured as the number of Web pages visited in session;
2. session duration (D), measured as the time interval (in seconds) between the first and the last user's clicks in session (this interval does not reflect the real time of a user—site interaction because the time for which the user browses the last page in session cannot be read from log);
3. mean time per page (M), which is the average time (in seconds) for which the user browsed a single page in session; it is determined according to the following formula:

$$M = \frac{D}{L-1} \tag{1}$$

   where D is the session duration and L is the session length, $L > 1$;
4. type of the source of the visit (S), depending on the page which had referred the user to the bookstore site. Six different types of sources were distinguished: references from organic search engine results, references from paid search engine results (such as Google Adwords ads), references from e-mail newsletters, entrances through social media sites (Facebook), internal references from other pages within the same website (outside the Web store), and other sources; $S \in \{$*paid search result*, *organic search result*, *newsletter*, *social media*, *internal reference*$\}$;
5. three binary variables corresponding to the facts of performing key operations related to the purchase transaction: *Register/Login* (R) connected with a user's registration or logging into the store site, *Adding a product to the shopping cart* (A), and *Buy* (B) corresponding to an order confirmation (i.e. successful finalization of purchase transaction);
6. a set of categories for products viewed in session.

Since user sessions reconstructed from log data are only an approximation of real user visits on a website, we used additional data gathered by SuperTracker (ST)

open source software for the osCommerce platform, to verify and refine session features. Information on categories for products viewed by customers was obtained by combining three data sources: log data (URIs of the requested content included in HTTP requests), ST data (numbers of categories viewed by customers), and the retailer's product database (the assignment of product numbers to categories).

4.4 Defining innovative and traditional customers

Various criteria may be adopted to divide customers into groups. In particular, one can apply unsupervised or supervised classification methods to divide a customer set or a session set into clusters (Carmona et al. 2012; Chang et al. 2007; Mohammadnezhad and Mahdavi 2012; Nenava and Choudhary 2013; Tanna and Ghodasara 2012). We decided to distinguish two customer groups: *innovative customers* and *traditional customers*, taking into consideration their preferences for types of products. In general, a distinction between customers preferring innovative or traditional products is not clear-cut because Web users may view, search, and add to shopping carts many different kinds of products in session. We defined two types of customers in the online bookstore in the following way:

1. *Innovative customers* are defined as users who during their interaction with the Web store site viewed some products which can be considered "innovative": audio-books and multimedia products (mainly films).
2. *Traditional customers* are defined as users who did not view any "innovative" products, but only typical products—printed books.

The proposed division into *innovative* and *traditional* customers is arbitrary and it was motivated by the experience of the online retailer who shared the data with us. The intuition for such a categorization was that users interested in audio-books and multimedia contents are mostly users who prefer multimedia-based leisure activities. Thus, they potentially cope better than other users with a virtual environment and characteristics of their sessions in a Web store will probably differ from those of users interested only in printed books. This was confirmed in our previous study (Suchacka and Chodak 2016) on the statistical analysis of sessions performed by *innovative* and *traditional* customers, defined in the same way as in this paper. The results showed that *innovative* customers open on average many more pages and spend much more time interacting with the e-commerce site than *traditional* customers and this tendency is even more visible if only buying sessions of both types are taken into consideration.

The division into *traditional* and *innovative* customers may be debatable but taking into consideration that multimedia products and printed books are two most popular categories in the analyzed bookstore, both in terms of the number of page views and the number of transactions, it seems to be justified. Finally, it is only an exemplification used in the experiment to present that the proposed approach can give different results for various customers groups. This division is arbitrary, but important for an online bookstore because of new trends concerning e-books and audio-books.

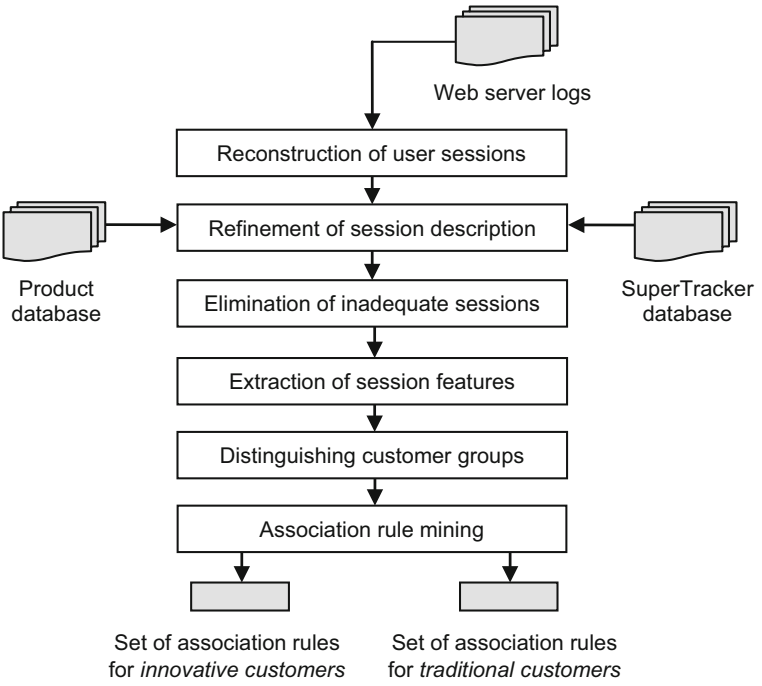

Fig. 1 Association rule generation framework for a B2C website

Each of the customer group (*innovative* vs. *traditional*) was separately analyzed using the approach described in Sect. 4.5. A methodology used to association rule mining is based on A-priori algorithm (Markov and Larose 2007). A scheme of our approach is presented in Fig. 1 and the notation used in the paper is summarized in Table 2.

4.5 Discovering association rules to assess purchase probability in a Web store

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of possible important events in a user session, distinguished taking key session features into consideration (see Sect. 4.3). I should be individually determined for the target Web store. In our case I contains the following 17 events:

1. $L = short$,
2. $L = medium$,
3. $L = long$,
4. $D = short$,
5. $D = medium$,
6. $D = long$,
7. $M = short$,
8. $M = medium$,

Table 2 List of the main symbols used in the paper

Symbol	Description
A	Session feature corresponding to adding a product to the shopping cart, $A \in \{true, false\}$
B	Session feature corresponding to an order confirmation (i.e. making a purchase), $B \in \{true, false\}$
C_k	Set of candidate k-element sets of events (potentially frequent event sets) in A-priori algorithm, $k = 1, 2, \ldots$
$Conf$	Confidence measure of the association rule strength (%)
D	Session feature corresponding to session duration (in seconds)
I	Set of possible events used to describe user sessions
L	Session feature corresponding to session length (in the number of pages)
L_k	Set of frequent k-element sets of events (those with minimum support Min_{Supp}) in A-priori algorithm, $k = 1, 2, \ldots$
M	Session feature corresponding to mean time per page (in seconds)
Min_{Conf}	Minimum confidence for strong association rules (%)
Min_{Supp}	Minimum support for strong association rules (%)
R	Session feature corresponding to a user's registration or logging into the site, $R \in \{true, false\}$
S	Session feature corresponding to a type of the source of the visit, $S \in \{paid\ search\ result, organic\ search\ result, newsletter, social\ media, internal\ reference\}$
$Supp$	Support measure of the association rule strength (%)
$Supp_{6+}$	Additional support measure of the association rule strength, based only on sessions in T with the length exceeding five pages (%)
T	Set of all analyzed user sessions

9. $M = long$,
10 $S = paid\ search\ result$,
11. $S = organic\ search\ result$,
12. $S = newsletter$,
13. $S = social\ media$,
14. $S = internal\ reference$,
15. $R = true$,
16. $A = true$,
17. $B = true$.

To be used in A-priori algorithm, numerical values of session length (L), session duration (D), and mean time per page (M) should be transformed into categorical variables by the discretization process. All three numerical variables were separated into three containers: *short*, *medium*, and *long*. Let us denote thresholds determining respective containers as L_1 and L_2 for the session length, D_1 and D_2 for the session duration, and M_1 and M_2 for the mean time per page.

Let T be a set of all analyzed user sessions in the observation window. Each session t in T is represented as a set of session events from I. For example, let us assume that for a given website the following threshold values had been set: $L_1 = 25$ pages, $L_2 = 45$ pages, $D_1 = 5$ min, $D_2 = 15$ min, $M_1 = 10$ s, and

$M_2 = 20$ s. Let us consider an example of a user who entered the site by following one of the links in an organic search engine's results, then logged on, searched for a few books, added one book to the shopping cart, and then finished the session; the session contained 14 pages and lasted 12 min, which results in the mean time per page equal to 55.4 s. Such a session will be represented as a 6-element set of events $\{S = organic\ \ search\ \ result,\ \ R = true,\ \ A = true,\ \ L = short,\ \ D = medium,\ M = long\}$.

Let us consider any k-element set of events $Z \subseteq I$ ($k = 1, 2, \ldots$). The frequency of the set of events Z is the number of sessions in set T which contain set Z. We say that Z is frequent if it is contained in at least Min_{Supp} percent of sessions belonging to T. A-priori algorithm used to determine association rules requires determining frequent k-element sets of events for different k.

The first stage of discovering association rules in our approach consists in finding all frequent sets of events. It is realized according to A-priori algorithm proposed in (Agrawal and Srikant 1994), the pseudocode of which is shown in Fig. 2.

Let C_k be a set of candidate k-element sets of events (i.e. potentially frequent sets of events) and L_k be a set of frequent k-element sets of events. Execution of A-priori algorithm starts from counting the support of 1-element event sets, C_1, and including in L_1 only the sets which are frequent. Unless the set of frequent $k-1$-element event sets is empty, two operations are performed iteratively. First, the frequent event sets L_{k-1} found in the $(k-1)$th pass are used to generate a set of candidate k-element event sets, C_k, using the apriori_gen function (C_k is generated by joining frequent event sets with $k-1$ elements and deleting those containing any subset that is not frequent). Next, the support of candidates in C_k is counted by checking all user sessions in T and only frequent k-element event sets are included in L_k. The final result of A-priori algorithm are all frequent event sets found in T, which are then used to generate association rules.

We apply association rules to sets of events describing user sessions in a Web store. Thus, an association rule has a form of the implication $X \Rightarrow Y$ (if X, then Y), where X and Y are event sets, $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \emptyset$. X is an *antecedent* of the rule and Y is a *consequent*. Our goal of applying A-priori algorithm to Web

Input: Set of user sessions T, minimum support Min_{Supp}
Output: All frequent event sets in T

1: $L_1 \leftarrow$ {frequent 1-element sets of events}
2: **for** ($k = 2$; $L_{k-1} \neq \emptyset$; k++) **do begin**
3: $C_k \leftarrow$ apriori_gen(L_{k-1});
4: **for all** sessions $t \in T$ **do begin**
5: $C_t \leftarrow$ subset(C_k, t);
6: **for all** candidates $c \in C_t$ **do**
7: c.count++;
8: **end**
9: $L_k \leftarrow \{c \in C_k \mid c.\text{count} \geq Min_{Supp}\}$;
10: **end**
11: Answer $\leftarrow \bigcup_k L_k$;

Fig. 2 Discovering frequent itemsets according to A-priori algorithm

session data is to identify these events in a user session which increase the probability of a product purchase. Thus, we aim to discover association rules of the form: $X \Rightarrow \{B = true\}$, where $X \subseteq I$ and $X \cap \{B = true\} = \emptyset$.

Each association rule is described with two measures: support and confidence, which have been typically used to identify strong rules (c.f. related work in Sect. 3).

The *support* for an association rule $X \Rightarrow Y$ is the percentage of sessions in T which contain both X and Y and it is determined as the number of sessions with X and Y divided by the overall number of sessions:

$$Supp = P(X \cap Y) \tag{2}$$

The *confidence* of an association rule $X \Rightarrow Y$ is the percentage of sessions in T containing X which also contain Y. It is determined as the number of sessions with X and Y divided by the number of sessions with X:

$$Conf = P(Y|X) = \frac{P(X \cap Y)}{P(X)} \tag{3}$$

An association rule is strong if it meets certain minimum confidence and support criteria. We denote the minimum support as Min_{Supp} and the minimum confidence as Min_{Conf}.

In addition to these two common measures we introduce an additional measure, $support_{6+}$, based only on sessions in T with the length exceeding five pages. In our case it is of a greater practical importance than the support measure because in reality a customer in this online bookstore has to visit at least six pages to finalize a purchase transaction. This is a consequence of the implementation of the checkout process in the software used in the analyzed bookstore (osCommerce platform). A customer enters the website via some page, e.g. a bookstore home page or any product description page. To make an order the customer has to visit the following five pages: (1) a create an account or login page, (2) a checkout shipping page with the choice of a delivery method, (3) a checkout payment page with the choice of a payment method, (4) a checkout confirmation page with a detailed information about the order displayed and a confirmation button, and (5) a checkout success page informing that the order was successfully placed. Usually buying sessions are much longer than six pages (Fig. 3) and include many product description pages. The minimum length of buying sessions in our set is 9 pages for *innovative customers* and 13 pages for *traditional* ones (Suchacka and Chodak 2016). However, it is possible that a customer had visited the bookstore before, had added products to the shopping cart during the previous visit, and starts a current buying session with a non-empty shopping cart. $Support_{6+}$ measure does not take "accidental", extremely short sessions into consideration and gives an idea of what percentage of sessions satisfying both the antecedent and consequent conditions one could predict in practice by observing the online behavior of customers.

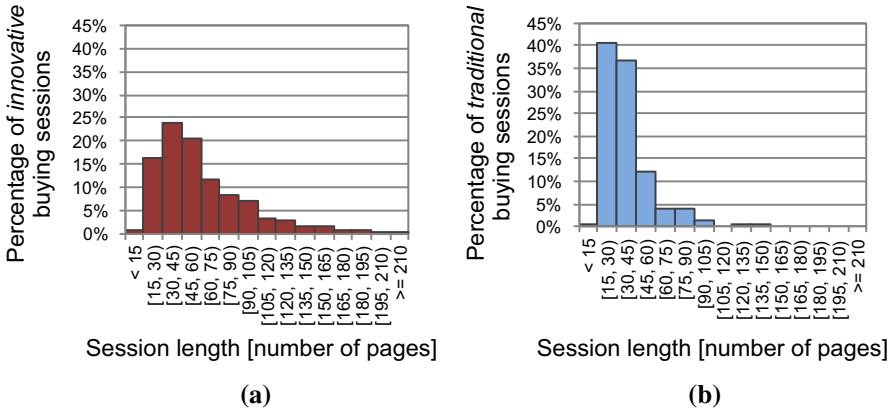

Fig. 3 Histogram of session lengths for buying sessions **a** for *innovative customers*; **b** for *traditional customers*. To easily compare the results for both customer groups histogram intervals have a fixed width and up to fifteen first intervals are shown in the figures (Suchacka and Chodak 2016)

5 Results

We used data recorded in server logs from April 01 to September 30, 2014 (13 552 772 HTTP requests, 4 GB in total). The user session set contained 33 354 user sessions. Among them 6 171 sessions were classified as performed by *innovative customers* (466 of these sessions, i.e. 7.55 %, ended with a purchase), and 5 415 sessions were performed by *traditional customers* (including 207 sessions, i.e. 3.82 %, ended with a purchase). Other sessions, which could not be clearly classified to any of these groups, were not taken into account in the analysis.

Parameters required for A-priori algorithm were determined experimentally. This group of parameters includes values of thresholds specifying *short*, *medium*, and *long* containers for the session length, the session duration, and the mean time per page to be used in association rule mining (L_1, L_2, D_1, D_2, M_1, and M_2, respectively). The experiments used results of our study characterizing sessions of *traditional* and *innovative customers* in the online bookstore (Suchacka and Chodak 2016). The following threshold values were used to build candidate association rules (in all possible combinations): 15 and 30 pages for L_1; 45, 60, and 75 pages for L_2; 300 and 600 s for D_1; 900, 1200, 1500, and 1800 s for D_2; 15 and 30 s for M_1; 30 and 45 s for M_2.

As we consider associations between session features in the context of a successful purchase transaction, and in reality a very small percentage of all sessions end at checkout, we assumed the minimum support, Min_{Supp}, equal to 1 % and the minimum confidence, Min_{Conf}, equal to 80 %. Top five association rules based on a confidence measure for *traditional* and *innovative customers* are shown in Tables 3 and 4, respectively.

Let us consider the strongest association rule for *traditional customer* sessions, denoted by R_{trad1} (Table 3). The rule has the following form: $\{L \in [30, 75), D \in [600, 1500), R = true\} \Rightarrow \{B = true\}$. The support value for this rule is 1.09 % which means that the rule applies to 1.09 % of *traditional customers*. It is not much

Table 3 Five strongest association rules determined for *traditional customer* sessions

Rule	Antecedent						Consequent	*Conf* (%)	*Supp* (%)	$Supp_{6+}$ (%)
	L (no. of pages)	*D* (s)	*M* (s)	*S*	*R*	*A*	*B*			
R_{trad1}	∈ [30, 75)	∈ [600, 1500)			*True*		*True*	92.19	1.09	2.51
R_{trad2}	∈ [30, 60)	∈ [600, 1500)			*True*		*True*	91.67	1.02	2.34
R_{trad3}	∈ [30, 75)	∈ [300, 1500)			*True*	*True*	*True*	91.14	1.33	3.07
R_{trad4}	∈ [30, 60)	∈ [300, 1500)	∈ [15, 45)		*True*		*True*	91.05	1.13	2.60
R_{trad5}	∈ [30, 75)		∈ [15, 45)		*True*	*True*	*True*	90.91	1.11	2.56

Table 4 Five strongest association rules determined for *innovative customer* sessions

Rule	Antecedent						Consequent	*Conf* (%)	*Supp* (%)	$Supp_{6+}$ (%)
	L (no. of pages)	*D* (s)	*M* (s)	*S*	*R*	*A*	*B*			
R_{inno1}	≥45			*Paid search result*	*True*	*True*	*True*	89.47	1.10	1.64
R_{inno2}	∈ [30, 75)			*Paid search result*	*True*	*True*	*True*	89.16	1.20	1.78
R_{inno3}		≥900	∈ [15, 45)	*Paid search result*	*True*		*True*	88.89	1.17	1.73
R_{inno4}	∈ [30, 75)		∈ [15, 45)	*Paid search result*	*True*		*True*	88.73	1.02	1.52
R_{inno5}	∈ [30, 75)	≥1500	∈ [15, 45)		*True*		*True*	88.57	1.01	1.49

but the maximum support value can be only 3.82 % (the percentage of *traditional* buyers). Taking only *traditional customer* sessions with at least six pages into consideration one can notice that the support of the R_{trad1} rule is much higher ($support_{6+}$ is equal to 2.51). The confidence of the rule equal to 92.19 % means that of all *traditional customer* sessions some sessions fulfilled the antecedent condition and as much as 92.19 % of these sessions ended with a checkout success.

Taking into consideration that for all users who view only printed books the probability of purchase is only 3.82 %, one can say that the rule R_{trad1} has significant support and very high $support_{6+}$. The rule R_{trad1} allows us to formulate the prediction that a logged user who has viewed only printed books, has been staying in the online store for 10–25 min, and opened from 30 to 75 pages, will decide to confirm the purchase with a probability of more than 92 %.

Comparing events in antecedents of the rules in Tables 3 and 4 one can observe that the associations discovered for the two customer groups differ from each other. The most significant factors increasing the probability of making a purchase by a *traditional customer* include such session characteristics as: the fact that the

customer logged on ($R = true$) and added some products to the shopping cart ($A = true$), the mean time per page in the range of 15–45 s, session length ranging from 30 to 75 pages, and the session duration ranging from 5 to 25 min. On the other hand, for an *innovative customer* the most important factors, in addition to being logged on and not having an empty shopping cart, include the session length exceeding 30 pages, the session duration exceeding 15 min, the mean time per page in the range of 15–45 s, and the fact that the user was referred to the store website by following a paid search engine link ($S = paid\ search\ result$).

The association rules for both groups have very high confidence, exceeding 88 % for *innovative customers* and 90 % for *traditional* ones. The support level exceeding 1 % is not very high, but one should notice that in reality very few sessions in the online bookstore end with a purchase (3.82 % of *traditional customers* sessions and 7.55 % of *innovative customer* sessions) and the maximum support value will never exceed the percentage of buying sessions in the analyzed set. Support values for five strongest association rules are similar for both customer groups, however taking into account that the percentage of buyers is much lower for *traditional customers* than for *innovative customers*, one can say that the support for this group is significantly better, in fact. A similar interpretation applies to $support_{6+}$, which is both objectively higher and relatively better for *traditional customer* sessions.

6 Discussion

Associations discovered for two customer groups are different. In the case of an *innovative customer*, i.e. a visitor viewing some "innovative" products, the highest probability of making a purchase (0.89) occurs when the customer was referred to the store website by following a paid search engine link, logged on, added some products to the shopping cart, and opened more than 45 pages. In the case of a *traditional customer*, i.e. a visitor viewing only typical products, the highest probability of a purchase (0.92) occurs when the customer logged on, opened from 30 to 75 pages, and has been staying in the store from 10 to 25 min.

The obtained confidence level for the strongest association rules is very promising and suggests the practical possible appliance of the research. Some doubts may be raised by a limited support being just over 1 %. One should remember, however, that the maximum support will never exceed the percentage of buyers, which is 3.82 % for *traditional customers* sessions and 7.55 % for *innovative* ones. Confidence is a more important measure for the online retailer because the ability to successfully predict a small percentage of buying sessions is a key to implementing a service strategy focused on most profitable customers. Such an approach is consistent with the Pareto principle, stating that a dominant part of a company's profit is usually generated by a relatively small group of the most profitable customers (Koch 2008) and the company should focus its attention and marketing efforts especially on that group. Furthermore, it is worth noting that association rules discovered in this study for two customer groups are much stronger than rules discovered for all bookstore visitors in (Suchacka and Chodak 2013), which were characterized by support of 0.5 % and confidence of about 0.7 %.

The proposed approach has some limitations. First of all, it is difficult to say to what extent our findings can be applied to other online stores. Association rules discovered for the analyzed website reflect its specificity and cannot be directly generalized to other online stores or even to other bookstores for many reasons. Many factors may influence customer behavioral and purchasing patterns on e-commerce websites, including the software used to implement the online store, the website structure, marketing techniques used by the company, its size, type of products in the store offer, and local or global nature of e-business. Furthermore, Web traffic patterns may differ depending on time and day of a customer visit or purchase. In many sectors seasonality factors should be taken into consideration, e.g. book consumption tends to be higher in the Christmas season and sales of tourist clothes and articles increases during the holidays.

Another open issue for discussion is how to divide customers into groups characterized by differentiated purchasing patterns. Our division into *traditional* and *innovative* customers is arbitrary, but important for the online bookstore because of new trends concerning e-books and audio-books. Such a division into *traditional* and *innovative* customers could be applied in many other retail sectors as well, e.g. in electronics sales into customers buying 2D TV sets and 3D TV sets, or in sales of clothes into customers looking for traditional fashions and those interested in latest fashion trends. For some e-commerce websites the division may be done automatically by applying various data mining techniques using information on customer online behavior. Distinguishing key customer groups may be performed based on established criteria or may result from the experience and observations of an online store manager.

Despite the above limitations our approach and the discovered association rules may suggest business implications for other online retailers. Due to a huge variety of implementation and organizational e-commerce solutions the approach cannot be automatically transplanted into other online stores but the proposed methodology may be adapted and followed. The ability to identify e-customers with a high probability of a successful purchase completion is very valuable for an online retailer as it makes it possible to focus on a group of key customers and organize the e-commerce service in the way that maximizes the number of orders placed online, the amount of sales and achieved revenue.

The possibility of identification of active sessions with a high probability of success is especially valuable for managers as it enables using some marketing techniques, like enhancements to the recommendation system dedicated only for such customers in order to increase the shopping cart value or to encourage customers to buy (e.g. via cross-selling, price promotions for complementary products, etc.). A large number of abandoned shopping carts is a significant problem on e-commerce sites (Rajamma et al. 2009). The observation of sessions with a high purchase probability allows an online retailer to quickly make contact with the customer in the case of an order not being completed to recognize a reason for a transaction failure. In the analyzed bookstore usually there is a reason for not completing the checkout process by a customer and it can be deduced from the customer's navigation path. If a customer had taken an effort to start the checkout process, i.e. they added some products to the shopping cart and gave their personal data during a registration process, it is very likely that there is some objective reason

why the order was not completed. In the case of the analyzed bookstore, the most frequent reasons are the following: (1) a customer is not sure of the actual delivery time and it is important for them to get products before a concrete date (e.g., birthday or other occasions); (2) a customer is not sure of some parameters of the product (due to incomplete or incorrect description on the Web page); (3) an institutional client is not sure if they can get the invoice, the invoice with a deferred payment, or invoices divided into parts due to the nature of financing (EU grants, municipal grants for schools, etc.). The bookseller confirms that a contact with such customers by phone or e-mail often results in the finalization of an order; customers' reactions are clearly positive, and such a contact is perceived as the concern for a customer. In practice it is impossible to analyze all abandoned shopping carts by the online retailer so knowledge of the rules helping to identify customers who are very likely to finish the checkout process would decrease the amount of hotline work.

Furthermore, the method for predicting online purchases may be used to develop and implement an intelligent server resource management policy. Under the server heavy load the system might assess purchase probability in active user sessions, assign priorities to them according to the assessed values, and enforce a priority-based admission control and scheduling algorithm instead of random rejection of client requests. Such a business-oriented service policy would allow for preventing the system overload and reducing the number of abandoned shopping carts and losses of potential revenue.

Discovery of differences in the factors increasing the purchase probability in *traditional* and *innovative customer* sessions has potential implications for the online retailer as it shows the sense of customers' service differentiation based on categories of products viewed by them. The bookseller can implement customized service strategies for visitors assigned to these groups based on the observation of their sessions' features. When different association rules are discovered for the specified customer groups, the marketing strategies targeted at these groups should be different. In our case, the finding that *innovative* buyers are often referred to the online bookstore by paid search engine results confirms the effectiveness of online advertising services used by the company with respect to audio-books and multimedia products and suggests improvements of this form of advertising applied to printed books.

7 Conclusion

In this paper we proposed an approach based on association rule discovery in e-customer sessions in order to assess the purchase probability in an online session. Two customer groups were distinguished based on customer preferences for types of products: *innovative customers* and *traditional customers*. The approach was applied to analyze historical data obtained from a real online bookstore.

Our results provide an important contribution to the identification of features characterizing online behavior of customers in online stores. New session features taken into consideration to build association rules include: mean time per page, type of the source of the visit, viewed products' categories, and the fact whether the user registered or logged into the site, whether they added a product to the shopping cart, and whether they placed an order. Different associations were discovered for two

customer groups and different session feature sets proved to be good purchase predictors. These findings confirm results of other studies showing that it is worth dividing customers due to some significant features of their navigational and purchasing behavior before the analysis.

The main advantage of our approach over other data mining techniques applied to purchase prediction is that it is relatively easy to implement on e-commerce websites and as a result it can easily adapt to changes in customer behavior patterns in real time. Although the discovered association rules cannot be generalized to other online stores, the proposed methodology may be easily adapted to the specificity of other e-commerce websites.

There are many possible future developments of our approach. First, more user session features can be utilized based of log files, e.g. an operating system, a client Web browser, time and date of purchase, or day of the week. The analysis of log data recorded in the longer period would make it possible to incorporate seasonality factors and a history of a customer's previous visits and purchases. Another direction of further research is a deeper analysis of abandoned shopping carts in the context of user navigation on the site in order to find behavioral patterns and possible reasons for giving up purchase intention and to identify sessions which are most likely to be abandoned before starting the checkout process.

In this paper two customer classes were distinguished a priori depending on the type of viewed products. In future work we are planning to examine various clustering methods to automatically divide the entire user session set into clusters taking information on product categories and behavioral patterns into consideration. We are also planning to compare the efficiency of association rule discovery and other data mining techniques in predicting online purchases. Such a comparative study could be performed using an e-commerce Web server simulator. Simulation experiments driven by a Web traffic model developed for real e-commerce data would make it possible to check how many user clicks in an active session must be observed to recognize a buyer.

Acknowledgments This work was partially supported by funds from the National Science Centre (NCN) in Poland under Grant No. 2013/11/B/HS4/01061.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Adnan M, Nagi M, Kianmehr K, Tahboub R, Ridley M, Rokne J (2011) Promoting where, when and what? An analysis of Web logs by integrating data mining and social network techniques to guide ecommerce business promotions. Soc Netw Anal Min 1(3):173–185

Adomavicius G, Tuzhilin A (2001) Expert-driven validation of rule-based user models in personalization applications. Data Min Knowl Disc 5(1–2):33–58
Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of the VLDB'94. Morgan Kaufmann Publishers, San Francisco, pp 487–499
Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In: Proceedings of SIGMOD'93. ACM, New York, pp 207–216
Borzemski L, Kamińska-Chuchmała A (2012) Client-perceived Web performance knowledge discovery through turning bands method. Cybernet Syst 43(4):354–368
Borzemski L, Suchacka G (2010) Business-oriented admission control and request scheduling for e-commerce websites. Cybernet Syst 41(8):592–609
Carmona CJ, Ramírez-Gallego S, Torres F, Bernal E, del Jesus MJ, García S (2012) Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. Expert Syst Appl 39(12):11243–11249
Catledge LD, Pitkow JE (1995) Characterizing browsing strategies in the World-Wide Web. Comput Netw ISDN 27(6):1065–1073
Chang H-J, Hung L-P, Ho C-L (2007) An anticipation model of potential customers' purchasing behavior based on clustering analysis and association rules analysis. Expert Syst Appl 32(3):753–764
Changchien SW, Lee C-F, Hsu Y-J (2004) On-line personalized sales promotion in electronic commerce. Expert Syst Appl 27(1):35–52
Chen Z, Fu AW-C, Tong FC-H (2004) Optimal algorithms for finding user access sessions from very large Web logs. World Wide Web 6:259–279
Chen Y-L, Kuo M-H, Wu S-Y, Tang K (2009) Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. Electron Commer Res Appl 8(5):241–251
Cheng C-H, Chen Y-S (2009) Classifying the segmentation of customer value via RFM model and RS theory. Expert Syst Appl 36(3):4176–4184
Chiang WY (2011) To mine association rules of customer values via a data mining procedure with improved model: an empirical case study. Expert Syst Appl 38(3):1716–1722
Cho YS, Moon SC, Oh I-B, Shin J-H, Ryu KH (2013) Incremental weighted mining based on RFM analysis for recommending prediction in u-commerce. Int J Smart Home 7(6):133–144
Chou P-H, Li P-H, Chen K-K, Wu M-J (2010) Integrating web mining and neural network for personalized e-commerce automatic service. Expert Syst Appl 37(4):2898–2910
Cooley R, Mobasher B, Srivastava J (1999) Data preparation for mining World Wide Web browsing patterns. Knowl Inf Syst 1:5–32
Deng X, Jin C, Higuchi Y, Han CJ (2010) An efficient association rule mining method for personalized recommendation in mobile e-commerce. In: Proceedings of ICEBI'10. Atlantis Press, Paris, pp 382–389
Han J, Pei J, Yin Y, Mao R (2004) Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min Knowl Disc 8(1):53–87
Hop W (2013) Web-shop order prediction using machine learning. Master Thesis, Erasmus University Rotterdam
Huiying Z, Wei L (2004) An intelligent algorithm of data pre-processing in Web usage mining. In: Proceedings of the IEEE WCICA'04, vol 4. New York, pp 3119–3123
Huk M, Kwiatkowski J, Konieczny D, Kędziora M, Mizera-Pietraszko J (2015) Context-sensitive text mining with fitness leveling genetic algorithm. In: Proceedings of the IEEE CYBCONF'15. New York, pp 183–188
Jenamani M, Mohapatra PKJ, Ghose S (2003) A stochastic model of e-customer behavior. Electron Commer Res Appl 2(1):81–94
Kazienko P (2008) Associations: discovery, analysis and applications. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław
Kim JK, Cho YH (2003) Using Web usage mining and SVD to improve e-commerce recommendation quality. In: Proceedings of PRIMA'03, LNCS 2891. Springer, Berlin, pp 86–97
Koch R (2008) The 80/20 principle: the secret of achieving more with less. Doubleday, New York
Kwan ISY, Fong J, Wong HK (2005) An e-customer behavior model with online analytical mining for internet marketing planning. Decis Support Syst 41(1):189–204
Lee Y-S, Yen S-J (2007) Mining Web transaction patterns in an electronic commerce environment. In: Proceedings of APWeb/WAIM'07 international workshops, LNCS 4537. Springer, Berlin, pp 74–85
Lee J, Podlaseck M, Schonberg E, Hoch R (2001) Visualization and analysis of clickstream data of online stores for understanding Web merchandising. Data Min Knowl Disc 5:59–84

Lee Y-C, Hong T-P, Lin W-Y (2005) Mining association rules with multiple minimum supports using maximum constraints. Int J Approx Reason 40(1–2):44–54
Markov Z, Larose DT (2007) Data mining the Web: uncovering patterns in Web content, structure, and usage. Wiley-Interscience, Hoboken
Mohammadnezhad M, Mahdavi M (2012) Providing a model for predicting tour sale in mobile e-tourism recommender systems. IJITCS 2(1):1–8
Morzy M (2006) Efficient mining of dissociation rules. In: Proceedings of DaWaK'06, LNCS 4081. Springer, Berlin, pp 228–237
Nenava S, Choudhary V (2013) Hybrid personalized recommendation approach for improving mobile e-commerce. IJCSET 4(5):546–552
Park JS, Chen M-S, Yu PS (1997) Using a hash-based method with transaction trimming for mining association rules. IEEE Trans Knowl Data Eng 9(5):813–825
Peng Y, Wan H (2010) An algorithm of commodities association rules mining in e-commerce based on rough set. In: Proceedings of IEEE iTAP'10. New York, pp 1–3
Poggi N, Moreno T, Berral JL, Gavald R, Torres J (2007) Web customer modeling for automated session prioritization on high traffic sites. In: Proceedings of UM'07, LNCS 4511. Springer, Berlin, pp 450–454
Rajamma RK, Paswan AK, Hossain MM (2009) Why do shoppers abandon shopping cart? Perceived waiting time, risk, and transaction inconvenience. J Prod Brand Manag 18(3):188–197
Sen A, Dacin PA, Pattichis C (2006) Current trends in Web data analysis. Commun ACM Entertain Netw 49(11):85–91
Shen Z-JM, Su X (2007) Customer behavior modeling in revenue management and auctions: a review and new research opportunities. Prod Oper Manag 16(6):713–728
Shim B, Choi K, Suh Y (2012) CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns. Expert Syst Appl 39(9):7736–7742
Stassopoulou A, Dikaiakos MD (2009) Web robot detection: a probabilistic reasoning approach. Comput Netw 53(3):265–278
Stevanovic D, Vlajic N, An A (2011) Unsupervised clustering of Web sessions to detect malicious and non-malicious website users. Procedia Comput Sci 5:123–131
Suchacka G (2014) Analysis of aggregated bot and human traffic on e-commerce site. In: Proceedings of IEEE FedCSIS'14, ACSIS, vol 2. New York, pp 1123–1130
Suchacka G, Borzemski L (2013) Web server support for e-customer loyalty through QoS differentiation. TCCI 12:89–107
Suchacka G, Chodak G (2013) Practical aspects of log file analysis for e-commerce. In: Proceedings of CN'13, CCIS 370. Springer, Berlin, pp 562–572
Suchacka G, Chodak G (2016) Characterizing Web sessions of e-customers interested in traditional and innovative products. In: Proceedings of ECMS'16. European Council for Modelling and Simulation, pp 635–640
Suchacka G, Skolimowska-Kulig M, Potempa A (2015a) A k-Nearest Neighbors method for classifying user sessions in e-commerce scenario. J Telecommun Inf Technol 3:64–69
Suchacka G, Skolimowska-Kulig M, Potempa A (2015b) Classification of e-customer sessions based on Support Vector Machine. In: Proceedings of ECMS'15. European Council for Modelling and Simulation, pp 594–600
Tanna P, Ghodasara Y (2012) Exploring the pattern of customer purchase with Web usage mining. In: Proceedings of ICAdC'12, AISC 174. Springer, New Delhi, pp 935–941
Thuan ND, Toan NG, Tuan NLV (2012) An approach mining cyclic association rules in e-commerce. In: Proceedings of IEEE NBiS'12. New York, pp 408–411
Totok A, Karamcheti V (2006) Improving performance of internet services through reward-driven request prioritization. In: Proceedings of IEEEIWQoS'06. New York, pp 60–71
Tsai C-F, Chen M-Y (2010) Variable selection by association rules for customer churn prediction of multimedia on demand. Expert Syst Appl 37(3):2006–2015
Tsay Y-J, Chiang J-Y (2005) CBAR: an efficient method for mining association rules. Knowl-Based Syst 18(2–3):99–105
Van den Poel D, Buckinx W (2005) Predicting online-purchasing behaviour. Eur J Oper Res 166(2):557–575
Wang Q, Makaroff DJ, Edwards HK (2004) Characterizing customer groups for an e-commerce website. In: Proceedings of EC'04. ACM, New York, pp 218–227

Wrzuszczak-Noga J, Borzemski L (2013) Applying the bidding mechanism in Web services with quality of service. In: Proceedings of CN'13, CCIS 370. Springer, Berlin, pp 582–591

Zatwarnicki K, Zatwarnicka A (2014) The cluster-based time-aware Web system. Proceedings of CN'14, CCIS 431. Springer, Berlin, pp 37–46

Zhang Y, Jiao JR (2007) An associative classification-based recommendation system for personalization in B2C e-commerce applications. Expert Syst Appl 33(2):357–367

Zhou X, Wei J, Xu C-Z (2006) Resource allocation for session-based two-dimensional service differentiation on e-commerce servers. IEEE Trans Parallel Distrib 17(8):838–850